**مقدمه**

**ایده ها و روش های پیشنهادی**

**معرفی استخراج دانش[[1]](#footnote-1) و شناسایی موجودیت های نامدار[[2]](#footnote-2)**

استخراج دانش و شناسایی موجودیت های نامدار دو روش برای بدست آوردن دانش ساخت یافته ساده از متن هستند. هدف از سیستم های استخراج دانش یافتن و درک بخش های مرتبط بهم در متن به صورت محدود است. بسیاری از این سیستم ها، به استخراج دانش از درون تکه های متن می­پردازند و یک نمایش ساختیافته از اطلاعات مرتبط بهم در متن را ارائه می­کنند. از اطلاعات بدست آمده در این سیستمها می­توان برای تولید پایگاه دانش بهره برد. این سیستم هادو هدف عمده را دنبال می­کنند:

* سازمان دهی اطلاعات داخل متن به نحوی کارا برای استفاده.
* قرار دادن اطلاعات در یک فرم دقیق منطقی که اجازه استنتاج و بررسی بیشتر را به الگوریتم ها و روش های کامپیوتری دیگر می­دهد.

بسیاری از الگوریتم های پیشرفته پردازش متن نمی­توانند با متن ورودی به صورت خام کار کنند و نیازمند دریافت برخی از اطلاعات به صورت ساختیافته هستند. مثلا به برچسب های مقوله دستوری[[3]](#footnote-3)یا اسامی موجود در متن یا نوع ارتباط آنها با هم در متن، نیازمندند.

سیستم های استخراج دانش معمولا دانشی که به صورت حقایق آشکار[[4]](#footnote-4) است را از متن بیرون می­کشند. اطلاعاتی نظیر «چه کسی در چه زمانی چه کاری را با چه کسی انجام داد؟».به عنوان نمونه درآمد یک نفر یا شرکت، دارایی های او، اعضای هیئت مدیره، دفاتر یک شرکت، دفتر مرکزی، مدیر عامل و استخراج هرگونه حقایق از این دست که از درون گزارش ها می­توان بیرون کشید. یا اسخراج کتاب های یک نویسنده، کارگردان یک فیلم سینمایی، محصولات یک کمپانی تولید خودرو یا نام مخترع یک وسیله با پردازش متون موجود در اینترنت. به عنوان نمونه به جمله زیر توجه کنید.

*اداره کل آموزش و پرورش و مراکز دفتری استانهای آن، در خیابان طالقانی تهران قرار دارند.*

ایده این است که سیستم بتواند جمله بالا را پردازش کرده و از داخل آن متوجه شود که «اداره کل آموزش و پرورش» و «مراکز دفتری استانها» نام دو سازمان است و به آنها برچسب ORGبدهد همچنین متوجه شود که این جمله درباره مراکز دفتری صحبت میکند و محل قرار گیری آنها را تعیین نموده است. بنابراین خروجی زیر را تولید کند:

|  |  |  |
| --- | --- | --- |
|  | Located(«اداره کل آموزش و پرورش», «خیابان طالقانی تهران»)  Located(«مراکز دفتری استانهای آموزش و پرورش», «خیابان طالقانی تهران») |  |

استخراج دانش همچنین در یافتن اطلاعات از متون پزشکی-زیست شناسی مانند روابط بین دارو های مختلف و بیماری ها، تاثیراتی ژنتیکی داروهای مختلف، تاثیر ژن در بیماری ها، اسامی گونه های مختلف گیاهان و ... مورد استفاده فراوانی قرار گرفته است.

اگر با دید کاربردی به استخراج دانش نگاه کنیم؛ این علم هم اکنون در بسیاری از تکنولوژی های امروزی موجود بوده و حتی فراگیرشده است. در بسیاری از کاربردها مانند سرویس ایمیل گوگل یا اپل یا اندیس گذاری اینترنتی. همانطور که در تصویر زیر مشاهده میکنید، سرویس ایمیل اپل به صورت اتوماتیک تاریخ و ساعت هایی که درون متن باشد را تشخیص می­دهد و این امکان را به کاربر می­دهدکه تنها با یک کلیک آن را در تقویم خود برای یادآوری ذخیره کند.



مثالی دیگر در جستجوی گوگل است. اگر به دنبال مقر اصلی یک شرکت چند ملیتی بگردیم؛ گوگل همراه با بازگرداندن نتایج جستجو، بهترین حدس خود را برای پاسخ به سوال ما، در ابتدا لیست نمایش می­دهد. به تصویر زیر توجه کنید.



یک زیر کلاس بسیار مهم از استخراج دانش، شناسایی موجودیت نامداراست که در آن، به دنبال یافتن و دسته بندی موجودیت های نامدار (اسامی) موجود در متن هستیم. بنابراین قدم اول یافتن اسامی است. در واقع محدوده اسامی در متن را تعیین می­کنیم.

*سرانجام در سومین روز هفدهمین دوره مسابقات آسیایی در اینچئون کره ‌جنوبی، طلسم بی ‌مدالی کاروان ایران شکسته شد و نجمه خدمتی، تیرانداز هجده ‌ساله تیم ملی کشور مان با درخشش خیره ‌کننده ‌ای به مقام قهرمانی آسیا رسید و نخستین طلای این رقابت‌ها را به نام کشورمان سند زد تا فراموش کنیم، ‌تیم فوتبال امید کشورمان امروز ‌رسما از رقابت‌های آسیایی اینچئون حذف شد .*

قدم بعدی دسته بندی اسامی یافته شده است.

*سرانجام در سومین روز هفدهمین دوره مسابقات آسیایی در اینچئون کره‌جنوبی، طلسم بی ‌مدالی کاروان ایران شکسته شد و نجمه خدمتی، تیرانداز هجده ‌ساله تیم ملی کشورمان با درخشش خیره ‌کننده‌ای به مقام قهرمانی آسیا رسید و نخستین طلای این رقابت‌ها را به نام کشورمان سند زد تا فراموش کنیم، ‌تیم فوتبال امید کشورمان امروز ‌رسما از رقابت‌های آسیایی اینچئون حذف شد.*

قرمز: سازمان –سبز: شهر –آبی: کشور –نارنجی: نام فرد–بنفش: قاره–سبز تیره: تاریخ

موجودیت[[5]](#footnote-5) به چیزی خیالی یا واقعی گفته می‌شود که به ذات خویش وجود دارد. به عنوان مثال هر انسان یک موجودیت است، دانشگاه شهید بهشتی، کشور ایران یا شرکت مخابرات ایران یک موجودیت هستند. اما هوا یا خاک موجودیت نیستند. ایده پشت موجودیت نامدار این است که برخی از موجودیت ها دارای نام هستند. مثلا من محمود راحت نام دارم، کشورم ایران نام دارد، شهرم تهران نام دارد اما صندلی که روی آن نشسته ام با این حال که یک موجودیت است؛ اما نام ندارد. البته بحث اینکه چه چیزهایی موجودیت نامدار هستند و چه چیزهایی نیستند، کمی فلسفی است و دیدگاه های مختلفی در این زمینه وجود دارد. اما در این بحث ما به دنبال اسامی و موجودیت های نامداری هستیم که به راحتی قابل تمیز باشند و از نظر فلسفی روی موجودیت بودن آنها شکی نیست. همچنین در پردازش متن چیزهایی مثل تاریخ (یازده خرداد سال 1364) و زمان (ساعت ده شب –11:30) و مقدار و نام مواد شیمیایی، پروتئنی یا زیستی نیز به عنوان موجودیت های نامدار در نظر گرفته می­شوند. با این حال که آنها به هیچ عنوان موجودیت نامدار نیستند ولی در زمان تعیین اسامی نامدار در متن عموما این اسامی نیز در نظر گرفته می­شوند.

در زیر برخی از کاربرد های شناسایی موجودیت های نامدار را مشاهده می­کنید:

* از موجودیت های نامدار در متن می­توان برای اندیس گذاری یا لینک دادن استفاده کرد. (بسیاری از شرکت ها از روشهای مختلف برای استخراج اسامی در متن استفاده می­کنند و در صفحات وب خود به آنها لینک می­زنند. مثلا در صفحات وب بایولوژی نام پروتئین ها و ژن های متن به صورت لینک بوده و به صفحه ای که به تبیین آن موجودیت می­پردازد لینک شده اند.)
* در کاربرد تشخیص نظر[[6]](#footnote-6) کاربران، نیازمندیم که بدانیم نظر نویسنده متن در مورد یک موضوع خاص مثبت یا منفی است. علاوه بر آن نیازمندیم بدانیم این نظر در مورد چه کسی یا چه چیزی است. لذا استخراج موجودیت های نامدار (شامل نام شرکت و نام محصول) می­تواند به تشخیص درست نظر کاربران کمک نماید.
* اگر به شناسایی موجودیت های نامدار به عنوان پیش نیاز استخراج دانش نگاه کنیم، کاربردهای زیادی برای آن وجود دارد. در استخراج دانش بیشتر کاری که انجام می­گیرد، استخراج موجودیت های نامدار و سپس تعیین روابط بین آنها است.
* کاربرد در سیستم های پرسش و پاسخ. در این سیستم ها پاسخ بسیاری از انواع سوالات، موجودیت های نامدارند. اینکه یک چیز چه زمانی اتفاق افتاده یا چه کسی آن را انجام داده است. بنابراین اگر یک سیستم خوب استخراج موجودیت نامدار برای زبان داشته باشیم، کمک شایانی به پاسخ سوالات به صورت اتوماتیک می­کند.

در ادامه برخی از سایت ها و پروژه های موجود در زمینه استخراج دانش برای زبان انگلیسی معرفی می­شوند. پروژه Open Calais از خبرگزاری Reuter’s یک نمونه جالب از استخراج اطلاعات است. در این سایت بخشی به شکل دموی آنلاین وجود دارد که می­توان یک متن انگلیسی دلخواه را در آن قرار داد. در شکل زیر نمونه خروجی آن را مشاهده میکنید.







همچنین با قرار دادن کرسر موس بر روی هر کدام از موجودیت های نامدار، دانش استخراج شده از متن برای آن نمایش داده میشود. شکل زیر.



پروژه AlchemyAPI برای 8 زبان انگلیسی، فرانسه، آلمانی، ایتالیایی، پرتغالی، روسی، اسپانیایی و سوئدی قابلیت استخراج موجودیت های نامدار را دارد. در شکل زیر نمونه ای از خروجی این سیستم را مشاهده می­کنید. متن ورودی، همان متن وارد شده در مثال بالا است.



از دیگر وب سایت ها میتوان به Evri و Yahoo’s term extraction اشاره نمود.

**ارزیابی سیستم های تشخیص موجودیت نامدار**

در این سیستمها، ورودی دنباله ای از کلمات و هدف شناسایی موجودیت نامدار است. برای اینکار باید شروع و پایان هر موجودیت و دسته مربوط به آن در متن تعیین گردد. به مثال زیر توجه کنید

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *رامین* | *مهمان* | *پرست،* | *سخنگوی* | *وزارت* | *امور* | *خارجه* | *دیروز* | *در* | *جمع* | *خبرنگاران* | *گفت* |
| PER | PER | PER | O | ORG | ORG | ORG | O | O | O | O | O |

به این مسئله می­توان به صورت یک مسئله دسته بندی کننده نگاه کرد که به هر کلمه ورودی یک برچسب می­زند. اما اینکار چندان درست نیست. زیرا در این مسئله ورودی دنباله ای از کلمات است و مثلا اگر یک کلمه با برچسب ORG داشته باشیم، احتمال اینکه کلمه بعدی نیز ORG باشد زیاد است. بنابراین روش ذکر شده همبستگی بین توکن ها در نظر نگرفته است.

استفاده از روش معمول دقت و فراخوان[[7]](#footnote-7) برای شناسایی موجودیت های نامدار چندان کارا نیست. به مثال زیر توجه کنید.

|  |  |  |  |
| --- | --- | --- | --- |
| *رامین* | *مهمان* | *پرست،* | *سخنگوی* |
| PER | PER | O | O |

در این مثال برچسب زننده[[8]](#footnote-8) ما بدنبال یافتن اسامی در متن است. به درستی دو توکن اول را برچسبPER زده است ولی توکن سوم را به اشتباه برچسب O زده است. در روش دقت و فراخوان چون کل عبارت «رامین مهمان پرست» یک نام است؛ اگر سیستم کل عبارات را درست تشخیص دهد، یکی به تعداد درست های تشخیص داده شده اضافه می­شود و در غیر این صورت یکی به تعداد غلط ها. حال در این شرایط علارغم اینکه سیستم جوابی نزدیک به درست داده و تنها بخشی از پاسخ را جا انداخته است؛ دو امتیاز منفی یکی برای اینکه پاسخ اش درست نبوده و دیگری برای اینکه یک نام را در متن نیافته است می­گیرد. این در حالیست که اگر سیستم کلا به همه توکن ها برچسب O میزد تنها یک خطا داشت. پر واضح است که این خلاف منطق است.

برای حل این مشکل پیشنهادات مختلفی برای ارزیابی سیستم های شناسایی موجودیت نامدار ارائه شده است. در این سیستم ها به برچسب زن، اعتبار جزئی برای حالاتی که موجودیت نامدار را به شکل تقریبا صحیح تشخیص داده است (مانند مثال بالا) می­دهند. یکی از این روشها امتیاز ماک[[9]](#footnote-9) است. البته چون محاسبه خود امتیاز ماک هم پیچیدگی های خاص خود را دارد؛ زیرا باید تعیین کنیم چگونه و به هر پاسخ نسبی چه امتیازی داده شود؛ بنابراین هنوز هم در بسیاری از منابع و مقالات از معیار اف یک[[10]](#footnote-10) برای گزارش نتایج سیستم خود استفاده می­کنند.

**استفاده از مدل توالی[[11]](#footnote-11) برای حل مسائل برچسب گذاری در استخراج دانش**

همانگونه که قبلا اشاره شد، برای حل یک مسئله برچسب گذاری در استخراج دانش میتوان به آن بصورت یک مسئله دسته بندی کلاسیک نگاه کرد. تک تک توکن های متن را استخراج نمود. آن ها را به صورت مستقل به دسته بندی کننده داد و برچسب آن را تعیین نمود. اما مشکل این روش در نظر نگرفتن همبستگی موجود در متن و ارتباط بین توکن های متوالی درون جمله است. با اینکار ویژگی توالی ذاتی کلمات در نظر گرفته نمیشود. به عنوان مثال در زبان انگلیسی تقریبا غیر ممکن است که یک فعل بعد از یک حرف تعریف در جمله بیاید. حال اگر در حال تعیین برچسب مقوله نحوی برای کلمات درون یک جمله انگلیسی باشیم، چنانچه از روش مطرح شده استفاده کنیم، چه بسا در مواردی به دو توکن متوالی برچسب حرف تعریف – فعل بزنیم که مطابق آنچه مطرح شد به احتمال زیاد اشتباه است. با در نظر گرفتن این مدل اطلاعات زیادی که در توالی توکن ها است از دست میرود. لذا به منظور بالا بردن دقت برچسب زننده خود، باید به نحوی اطلاعات توکن های مجاور را در تصمیم گیری خود منظور کنیم.

در ادامه بحث توضیحات را روی مسئله شناسایی موجودیت نامدار که یک مسئله کلاسیک در زمینه برچسب گذاری توالی در استخراج دانش است، متمرکز میکنم. برای استفاده از مدل توالی در شناسایی موجودیت نامدار نیازمند داده های آموزشی با ناظر هستیم. یعنی مجموعه از متون آموزشی داریم که در آنها کلمات بر اساس موجودیت شان برچسب صحیح خورده اند. بدین ترتیب مراحل زیر را برای آموزش سیستم داریم:

* یک مجموعه متن که نماینده متون آزمایشی ما هستند و در آنها موجودیت هایی که به دنبال یادگیری آنها هستیم وجود دارد.
* در این مرحله بر روی کلمات متون حرکت می­کنیم. برچسب موجودیت صحیح را به آن می­دهیم و اگر شامل هیچکدام از موجودیت های ما نمی­شد، برچسب O که نماینده Other است را به آن میدهیم.
* در بخش یادگیری ماشین، مدلی برای استخراج ویژگی که مناسب برای متن و برچسب های هدف باشد طراحی میکنیم.
* آموزش یک دسته بندی کننده توالی برای تعیین برچسب های هدف در متن.

در این مرحله مدل لازم برای برچسب زنی داده های آموزشی تولید شده است. به منظور آزمایش مدل مراحل زیر را دنبال میکنیم:

* جمع آوری مجموعه ای از متون آزمایشی
* اعمال مدل یادگیری شده برای استنتاج برچسب مناسب برای کلمات درون متن. این مدل به ما احتمال هر برچسب را به ازای کلمه ورودی میدهد.
* در خروجی برچسب مناسب را بر اساس احتمالات محاسبه شده تولید می­کنیم.

برای فهم بهتر مطلب یک مثال می­زنیم. در مثال زیر دو روش برای برچسب گذاری مشاهده می­شود.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| painting | new | ‘s | Picasso | Pablo | Sue | showed | Fred |  |
| O | O | O | PER | PER | PER | O | PER | IO Labeling |
| O | O | O | I-PER | B-PER | B-PER | O | B-PER | IOB Labeling |

در روش اول (IO Labeling) به هر کلمه در صورتیکه موجودیت داشته باشد برچسب آن را میدهیم و در غیر این صورت برچسب O یعنی بدون موجودیت میدهیم. اما مشکلی در این برچسب گذاری وجود دارد! در جمله بالا سه نام Fred، Sue و Pablo Picasso وجود دارد. اما برچسب گذاری ما تنها دو نام را نشان میدهد. در واقع هیچ روشی برای اینکه نشان دهیم در یک دنباله برچسب، نام دو نفر وجود دارد نداریم. برای حل این مشکل می­توان از روش دوم استفاده کرد.

در روش دوم (IOB Labeling) در صورتیکه یک موجودیت در دنباله شروع شود به آن برچسب B-Entity میدهیم. مثلا به شروع نام یک فرد مانند Pablo برچسب B-PER میدهیم (B ابتدای کلمه Begin است). و در صورتیکه کلمه ادامه موجودیت قبلی باشد به آن برچسب I-Entity میزنیم. مثلا به کلمه Piccaso چون ادامه نام است برچسب I-PER میزنیم (I ابتدای کلمه Inside است). مشابه روش قبل در صورتیکه کلمه موجودیت نباشد به آن برچشب O میدهیم. با این روش مشکل روش اول حل میشود. اما برچسب گذاری دوم از نظر هزینه محاسباتی مشکلاتی دارد. در روش اول اگر فرض کنیم c موجودیت داریم. در مجموع برچسب خواهیم داشت (به اضافه یک برای برچشب O است). اما در روش دوم برچسب خواهیم داشت (ضریب 2 برای شروع و پایان ها است). در نگاه اول این تفاوت ناچیز به نظر می­رسد. اما توجه کنید که اگر از مدل توالی برای برچسب زنی استفاده کنیم باید حداقل مجموعه حالات را به توان مرتبه مدل توالی برسانیم یعنی حداقل به توان 2. بنابراین برای روش اول حالت و برای روش دوم خواهیم داشت. بنابراین در نهایت استفاده از روش دوم موجب پایین آمدن شدید سرعت اجرا میگردد. با این حال که روش دوم بر اساس برخی نقطه نظرات روش درست برای برچسب زنی است؛ بسیاری از مقالات نیز از همین روش استفاده میکنند؛ اما به سه دلیل ما روش اول را برای برچسب زنی برمی­گزینیم.

1. بسیار سریع تر است. آموزش آن با داده کمتر و در زمان کوتاه تری امکان پذیر است.
2. مشکل فقط زمانی رخ میدهد که ما دو موجودیت با کلاس یکسان را به صورت متوالی در ورودی داریم. مثلا نام دو نفر پشت هم آمده است. این حالت در زبان انگلیسی به ندرت رخ می­دهد. در زبان فارسی هم که هدف اصلی ما است، این حالات به خاطر وجود علامات و کلمات ربط مانند «و» و «،» یا وجود کلمه «را» بعد از مفعول، کمتر از زبان انگلیسی رخ می­دهد.
3. سیستم هایی که از روش دوم استفاده میکنند، در عمل به ندرت می­توانند در خروجی پایان یک موجودیت و شروع همان نوع موجودیت را بدرستی تشخیص دهند و عموما برای جمله مورد مثال خروجی زیر را تولید می­کنند.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| painting | new | ‘s | Picasso | Pablo | Sue | showed | Fred |  |
| O | O | O | I-PER | I-PER | B-PER | O | B-PER | IOB Labeling in real practice |

که این نوع دسته بندی دقیقا معادل روش اول است و مزیتی نسبت به آن ندارد. لذا در عمل استفاده از روش اول علارغم اندکی مشکل منطقی تر است.

**ویژگی های مورد استفاده در روش مدل توالی**

اولین ویژگی واضح خود کلمات هستند. برای هر کلمه یک ویژگی برای بررسی احتمال قرار گرفتن آن کلمه در کلاس مورد نظر داریم. لذا این ویژگی در واقع مانند یک لغت نامه یادگیری شده از کلمات برای هر کلاس عمل می­کند. همچنین ویژگی هایی برای کلمات قبل و بعد از کلمه در حال پردازش قرار می­دهیم. مثلا میدانیم که بعد از کلماتی مانند «در» یا «به» احتمال آمدن یک موجودیت مکان با برچسب LOC بیشتر است. (علی به مدرسه رفت. حسن در خیابان انقلاب ماند.) چنانچه نتایج پردازش های زبانی دیگر مانند برچسب مقوله نحوی را داشته باشیم، بهتر است از آنها هم به عنوان ویژگی استفاده کنیم. برچسب مقوله نحوی کلمه در حال پردازش، کلمه قبل و بعد از آن مورد استفاده قرار می­گیرند.

تمام این ویژگیها تنها به داده ورودی توجه می­کنند و می­توان با آنها یک دسته بندی کننده معمولی بدون مدل توالی ساخت. تنها زمانی دسته بندی کننده ما در ساختار مدل توالی قرار می­گیرد که اطلاعات برچسب های کلمات مجاور را نیز در تعیین برچسب کلمه جاری مورد استفاده قرار دهد. به عنوان مثال وقتی در حال پردازش جمله «نام او شهید ابراهیم همت بود.» هستیم. اگر دسته بندی کننده توالی ما متوجه شود که توکن «ابراهیم» برچسب نام فرد (PER) دارد، بنابراین با احتمال زیاد برچسب توکن بعدی «همت» نیز PER خواهد بود. زیرا نام افراد معمولا بیش از یک توکن درازا دارد. و به این صورت میتوان ویژگی هایی داشت که این یک دنباله را مدل کنند.

یک دسته دیگر از ویژگی ها که در شناسایی موجودیت نامدار مورد استفاده اند؛ ویژگی حروف سازنده زیر رشته[[12]](#footnote-12) است. این ویژگی ها برای دسته بندی کلمات میتوانند بسیار کارآمد باشند. روش کار به این صورت است که یک زیر رشته از حروف را به عنوان ویژگی در نظر می­گیریم. مثلا اگر میخواهیم اسامی داروهای به زبان انگلیسی را استخراج کنیم، میتوان زیر رشته «oxa» را به عنوان ویژگی در نظر گرفت. در مثال زیر چند ویژگی حروف سازنده زیر رشته برای کاربرد تشخیص 5 دسته موجودیت شامل اسامی داروها، شرکت ها، فیلم ها، مکانها و اشخاص را مشاهده می­کنید.

 همانطور که در شکل میبینید، ویژگی زیر رشته ای «oxa» با دقت 100 درصد فقط در اسامی داروها در داده آموزشی دیده شده است. و برای دسته بندی کردن بسیار مناسب است. از طرف دیگر ویژگی زیر رشته ای «:» اغلب در اسامی فیلمها به زبان انگلیسی دیده شده است. زیر رشته «field» نیز در اسامی اشخاص، فیلم ها و مکان ها بیشتر دیده شده است.

ویژگی دیگری که کاملا مکمل ویژگی کاراکترهای سازنده زیر رشته است، دنباله شکل کلمات[[13]](#footnote-13) است. ایده این روش اولین بار توسط مایکل کالینیز [?] مطرح شد. در این روش کلمات را بر اساس شکل ظاهریشان به یکسری کلاس نگاشت میکنیم. این کلاس ها نمایشی ساده از کلمات اند که خواصی از آنها مانند طول کلمه، بزرگ و کوچک بودن آن (برای انگلیسی)، استفاده از اعداد یا علائم نگارشی درون آن را در خود دارند. شکل زیر نمونه ای از این کلاس ها را برای زبان انگلیسی ارائه می­کند.



در این سیستم استخراج ویژگی؛ هر حرف بزرگ به X، هر حرف کوچک به x، هر عدد به d و علائم نگارشی مانند -:, به خودشان نگاشت می­شوند. علاوه بر این در تولید کلاس ها نکته دیگری نیز مد نظر قرار گرفته است. ایده این است که از هر کلمه تنها دو حرف شروع و پایانی آن مورد استفاده قرار بگیرد. بنابراین اگر کلمه کمتر یا مساوی 4 حرف داشته باشد؛ کل آن در کلاس نماینده اش نگاشت میشود، در غیر این صورت برای تمام حروف بین دو حرف اول و آخر کلمه، فقط نوع حروف را در نمایش کلاسی میآوریم. در مثال بالا برای محاسبه کلاس کلمه «Varicella-zoster» به ازای Va ابتدای کلمه Xx، به ازای er انتهای کلمه xx و به ازای «ricella-zost» تنها x- که نشاندهنده نوع حروف در این زیر رشته است در کلاس نماینده شان آمده است که مجموعا موجب میشود تا این کلمه به کلاس Xx-xxx نگاشت شود. روشهای زیاد دیگری نیز وجود دارند که میتوان این کلاس ها را تعریف کرد. ولی نکته اصلی این است که این کلاسها اطلاعات کلی از ظاهر و حروف سازنده کلمه را در خود نهفته اند. بدین ترتیب کلمات را فضای بسیار کوچکتر و چگالتری میتوان نگاشت کرد.

**مدل توالی بیشینه عدم قطعیت[[14]](#footnote-14)**

داده ها در اغلب مسائل پردازش متن به صورت توالی از حروف، کلمات، عبارات، خطوط و جملات هستند و کار ما در بسیاری از موارد برچسب زدن به این توالی رشته ای است. یک نمونه بارز از این مسئله، برچسب مقوله نحوی است. نمونه های دیگر شامل شناسایی موجودیت نامدار، برچسب نقش معنایی[[15]](#footnote-15)، شناسایی انتهای جملات، قطعه بندی[[16]](#footnote-16) و تشخیص انتهای کلمه در زبان چینی است. در مدل توالی بیشینه عدم قطعیت، که به آن مدل مارکوف بیشینه عدم قطعیت[[17]](#footnote-17) نیز گویند، دسته بندی کننده به ازای هر توکن ورودی یک تصمیم میگیرد. این تصمیم مشروط بر شواهد حاصل از مشاهدات و تصمیمات قبلی است. مثال زیر کاربرد این روش در برچسب گذاری مقوله نحوی را نمایش می­دهد. برچسب سه توکن اول تعیین شده اند.



در نقطه تصمیم گیری ویژگی های مورد استفاده برای دسته بندی کننده شامل کلمه جاری، کلمه بعدی، کلمه قبلی، برچسب کلمه قبلی، برچسب دو کلمه قبلی به صورت متصل مورد استفاده است. همچنین سایر ویژگی های مشابه آنچه قبلا بررسی شد نیز کاربرد دارند. ایده به کار رفته در دسته بندی نیز مشابه دسته بندی کننده ماکزیمم عدم قطعیت است که پیشتر مورد بررسی قرار گرفت. شکل زیر کلیات نحوه حرکت از دسته بندی کننده ابتدایی ماکزیمم عدم قطعیت به مدل توالی بیشینه عدم قطعیت را نشان می­دهد.



در این شکل داده های توالی در بالا سمت چپ قرار دارند. هدف دسته بندی در مرحله توالی است. همانطور که در بالا راست دیده میشود، کلمات را در ورودی داریم و میخواهیم به هر کدام برچسب مناسب بزنیم. برای اینکار ما به هر تصیمیم به صورت مستقل نگاه میکنیم. به ازای هر تصمیم مجموعه از داده ها به صورت محلی (پایین سمت چپ) وجود دارد. این داده ها شامل کلمه جاری، کلمه قبلی، برچسب کلمه قبلی و ... است. سپس بر روی آن داده ها به استخراج ویژگی می­پردازیم. حال ما ویژگی های محلی را داریم، برچسب ها را هم بخاطر استفاده از یادگیری با ناظر داریم. با استفاده از آنها یک مدل بیشینه عدم قطعیت می­سازیم. در این مرحله تمام مسائلی که قبلا مطرح شده بود شامل بهینه سازی و هموارسازی های لازم انجام می­گیرد. در انتها یک دسته بندی کننده داریم.

چنانچه در هر مرحله برچسب توکن تعیین گردد و به مرحله بعد برویم، در واقع از یک روش حریصانه[[18]](#footnote-18) برای تعیین برچسب ها استفاده کرده ایم. اما در بسیاری از موارد بدنبال بهترین دنباله برچسب هستیم و علاقه مندیم تا فضای حالات برچسب ها را بهتر جستجو کنیم. زیرا مثلا فرض کنید به توکن اول برچسب A را داده ایم. بعد یک توکن دیگر را پردازش کرده ایم و به توکن سوم رسیدیم. در این مرحله بر اساس برخی شواهد متوجه میشویم که برچسبی که به توکن اول داده ایم نادرست بوده و به تبع آن برچسب توکن دوم نیز غلط است. حال ناگزیریم دوباره به توکن اول بازگشته، تصمیم خود را اصلاح کنیم و کار را از ابتدا تکرار کنیم. برای حل این مشکل چند روش وجود دارد. در ادامه با دو روش آشنا می­شویم.

**استنتاج پرتویی[[19]](#footnote-19)**

اولین این روش های استنتاج پرتویی است. در این روش به جای اینکه تنها بهترین جواب را نگهداریم، k تا از بهترین دنباله نتایج را ذخیره میکنیم. هر توالی از پاسخ ها را جداگانه ادامه میدهیم. بنابراین برای هر توکن ورودی باید k بار استنتاج انجام دهیم. این روش بسیار سریع بوده و پیاده سازی آن نیز آسان است. عموما اندازه k برابر با 3 یا 5 است. این اندازه برای استنتاج پرتویی نشان داده است که در بیشتر موارد میتواند بهترین توالی برچسب ها را محاسبه کند. ولی در بسیاری از موارد هم ممکن است که بهترین دنباله از برچسب ها از مجموعه پاسخ های استنتاج پرتویی خارج انداخته شود و در نتیجه جواب حاصل بهترین پاسخ توالی بر چسب نباشد. لذا در این روش هیچ تضمینی برای تولید بهترین پاسخ وجود ندارد.

**استنتاج ویتربی[[20]](#footnote-20)**

این روش بهتر از استنتاج پرتویی عمل میکند. در این روش بهترین توالی برچسب ها محاسبه می­گردد. به آن در پردازش متن استنتاج ویتربی یا الگوریتم ویتربی گفته می­شود. زیرا آقای اندرو ویتربی الگوریتم های زیادی برای حل این مشکل ارائه کرده است. پایه این روش بر برنامه نویسی پویا[[21]](#footnote-21) است. در آن یک پنجره کوچک تاثیر گذاری در نظر می­گیریم. برای انجام هر تصمیم گیری، پارامترهای تاثیر گذار تنها درون پنجره قرار دارند و هرچیزی خارج از آن را در نظر نمیگیریم. در این شرایط می­توان با استفاده از الگوریتم های برنامه نویسی پویا بهترین توالی برچسب را محاسبه نمود. مزیت آن در تضمین ارائه بهترین پاسخ است. مشکلات آن هم در سختی پیاده سازی و مجبور کردن کاربر به استفاده از برچسب های توکن های قبلی فقط در محدوده پنجره تعریف شده است.

**میدان تصادفی شرطی[[22]](#footnote-22)**

میدان تصادفی شرطی یکی دیگر از روشهای برچسب گذاری احتمالاتی توالی است. با رابطه زیر ارائه میگردد.

احتمالات c و d روی کل دنباله تعریف می­شوند و مربوط به یک نقطه خاص در فضای مسئله نیستند. در نظر اول اینکار بسیار مشکل به نظر میرسد زیرا فضای حالات c و d نمایی است. اما به شرط حفظ ویژگی های به صورت محلی می­توان جوری برنامه نویسی پویا را انجام داد که تابع درست نمایی شرطی[[23]](#footnote-23) دنباله را به درستی محاسبه نمود. آموزش مطمئنا نسبت به سایر روشهای استنتاج کند تر است ولی از نظر تئوری برتری هایی نسبت به مدل مارکوف بیشینه عدم قطعیت دارد. این روش در بسیار از مقالات به عنوان لبه تکنولوژی[[24]](#footnote-24) برای برچسب گذاری توالی مورد استفاده است. اما لازم به ذکر است که در عمل، نتایج حاصل از این روش نزدیک به نتایج مدل مارکوف بیشینه عدم قطعیت است. در چنین حالاتی چیزیکه تاثیر بیشتری دارد، انتخاب ویژگی های قدرتمند و مناسب است. و استفاده از مدل مارکوف بیشینه عدم قطعیت نیز برای برچسب گذاری توالی کاملا منطقی و قابل قبول است.

**مدل بیشینه عدم قطعیت و تخمین ممیزی[[25]](#footnote-25)**

تاکنون مدل های مولدی[[26]](#footnote-26) زیادی مانند دسته بندی کننده نایو بیز[[27]](#footnote-27) و مدل زبانی[[28]](#footnote-28) معرفی شده اند. ویژگی این مدل ها در این است که همگی توانایی تولید (سنتز) داده طبق مدل آموزش یافته را دارند. اما امروزه کاربردهای زیادی برای مدل های شرطی ممیزی[[29]](#footnote-29) در پردازش متن، صوت، استخراج دانش و به طور کلی یادگیری ماشین دیده می­شود. ویژگی مشترک این مدل توانایی دسته بندی داده ها بر اساس مدل آموزش داده شده است. این رویکرد جدید به دلایل زیر صورت گرفته است

* کارایی و دقت بالا در دسته بندی دارند.
* می­توان به سادگی اطلاعات زبانی زیادی را به کمک این روش ها، مدل کرد.
* در تولید مدل های زبانی قابل استفاده مجدد کارآمدند.

در ابتدا تفاوت های مدل تولیدی و مدل تمایزی را بررسی میکنیم. اهمیت دانستن تفاوت های این دو مدل در انتخاب مدل مناسب برای حل مسئله است. دیده شده است که در بسیاری از موارد محققین به اشتباه برای حل یک مسئله دسته بندی از مدل های مولدی و به عکس برای حل مسئله سنتز داده از مدل های ممیزی استفاده کرده اند. البته نمیتوان به کلی انجام اینکار را رد کرد ولی بهتر است برای رسیدن به نتایج مناسب در انتخاب مدل دقت کرد. در هر دو دسته مدل، فرض میکنیم مجموعه داده به صورت داریم که در آنها d داده مشاهده شده و c کلاس داده که به صورت مخفی است. در مدل های مولدی (توزیع توام[[30]](#footnote-30))، احتمالات را روی جفت داده و کلاس مخفی به فرم قرار میدهند. (احتمال به صورت توزیع توام در نظر گرفته می­شود.) در واقع داده مشاهده شده را از روی اطلاعات مخفی تولید می­کنند. این روش ها تمام تکنیکهای کلاسیک پردازش متن شامل مدل چندگرمی[[31]](#footnote-31)، نایو بیز، مدل مخفی مارکوف[[32]](#footnote-32) و ... را در بر می­گیرند. هر جا تابع توزیع احتمال توام داده و کلاس را داشته باشیم، میتوانیم از این توزیع برای سنتز داده مشابه آنچه آموزش دیده ایم استفاده کنیم.

در نقطه مقابل، روشهای ممیزی (شرطی)، به شکل مستقیم تری دسته بندی که می­خواهیم انجام دهیم را مورد هدف قرار می­دهند. در واقع تابع احتمال را به صورت شرطی می­بینند. احتمال کلاس به شرط داده مشاهده شده . روش های ممیزی شامل رگرسیون لاجستیک[[33]](#footnote-33)، مدل بیشینه آنتروپی و میدان تصادفی شرطی می­شوند. البته بسیاری از دیگر روشهای یادگیری ماشین مانند ماشین بردار پشتیبان[[34]](#footnote-34) و شبکه عصبی نیز در دسته روشهای ممیزی برای دسته بندی قرار می­گیرند که در واقع به صورت مستقیم بر مبنای احتمالات قرار نگرفته اند.

یک روش برای نمایش تفاوت بین دو مدل، استفاده از مدلهای گرافی احتمال[[35]](#footnote-35) است. یک مدل گرافی، مدلی احتمالاتی است که در آن استقلال شرطی بین متغیرها توسط یک گراف نشان‌داده می‌شود. دو نوع مدل گرافی وجود دارد: مدل با گراف جهت‌دار که شبکه بیزی نامیده می‌شود و مدل با گراف ساده که میدان تصادفی مارکفی نامیده می‌شود (Bishop, 2006) (Daphne Koller, 2009). در مدل گرافی، متغییرها را با دایره و خطوط بین دایره ها نشاندهنده وابستگی مستقیم بین متغییرها است. برخی از متغییرها به صورت معمولی قابل مشاهده اند و برخی مخفی هستند. در شکل زیر، مدل گرافی احتمالی برای دو دسته بندی کننده رگرسیون لاجستیک (به نمایندگی از مدل ممیزی) و نایو بیز (به نمایندگی از مدل مولدی) نمایش داده شده است.



در نایو بیز در زمان دسته بندی، کلمات مختلف متنی را که به ما داده شده است داریم (d1, d2,d3). می­خواهیم بر اساس آن کلاس داده (c) را تعیین کنیم. از دید احتمالاتی، ما احتمال اولیه کلاس و احتمال هر کلمه به شرط کلاس را داریم. جهت فلش از کلاس به داده (مولدی) است که نشان می­دهد که کلمات از کلاس تولید شده­اند. و در واقع چیزی را که واقعا مشاهده کرده ایم را پیش گویی نمی­کند. اما وضعیت در رگرسیون لاجستیک معکوس این حالت است. دوباره ما کلامات درون متن را مشاهده می­کنیم و به دنبال پیشگویی کلاس داده هستیم. اما این بار ما به شکل مستقیم یک احتمال را بر روی کلاس به شرط داده مشاهده شده قرار میدهیم .

بنابراین در مدلهای مولدی به دنبال توزیع توام احتمال داده و کلاس و بیشینه کردن درستنمایی احتمال هستیم. انجام اینکار همانطور که می­دانیم بدیهی است. انتخاب وزنها برای این مدل بر اساس شمارش وقوع رخدادهای مختلف یک چیز و سپس تقسیم کردن بر یک مخرج نرمال کننده انجام می­گیرد. این روش برای تخمین بسیار کاربردی است.

در نقطه مقابل، در مدل شرطی ممیزی، میخواهیم احتمال را محاسبه کنیم. در این مدل سعی می­کنیم به صورت مستقیم، احتمال درستنمایی شرطی را بیشینه کنیم. احتمال کلاس های مشاهده شده به شرط داده. در ادامه خواهیم دید که اینکار به مراتب مشکلتر بوده ولی در عوض مفیدتر است زیرا به شکل مستقیم به درستی و غلطی دسته بندی کننده مربوط شده و نتایج بهتری را تولید خواهد کرد.

برای بررسی بهتر این قضیه به مقاله ای در زمینه مقایسه مدل های مولدی و ممیزی در زمینه رفع ابهام معنای کلمه[[36]](#footnote-36) توجه کنید (Dan Klein, 2002). در این مقاله، هدف تعیین معنای صحیح کلمه از بین معانی مختلف آن در متن است. به عنوان مثال کلمه «شیر» در فارسی می­تواند معانی بسیار متفاوتی داشته باشد. شیر خوردنی، شیر سلطان جنگل، شیر آب و شیر به معنای عصاره هر چیزی. در این مقاله سعی می­شود با دیدن کلمه شیر درون متن (با کمک شواهد در اطراف کلمه) تشخیص دهیم کدامیک از این معانی مد نظر نویسنده بوده است. دو مدل که از هر نظر کاملا با هم یکسان اند برای حل مسئله در نظر گرفته شده است. هر دو فیچر های یکسان، روش هموارسازی یکسان[[37]](#footnote-37) و داده های آموزشی و آزمایشی یکسانی دارند. تنها تفاوت دو روش حل مسئله، در این بوده که در یکی از روش مولدی و دیگری ممیزی برای دسته بندی استفاده شده است. در جدول زیر نتایج حاصل از تحقیق را مشاهده می­کنید.



اگر به نتایج حاصل روی داده های آزمایشی توجه کنید می­بینید که روش ممیزی کارایی بالاتری نسبت به مدل مولدی دارد (حدود 2.5 درصد). در این مثال روش مولدی مورد استفاده نایو بیز است. نکته دیگر این است که با توجه به نتایج داده های آموزشی متوجه می­شویم که روش تمایزی حدود 12 درصد کارایی بالاتری دارد. ممکن است در نظر اول این نتیجه ما رو خوشحال کند اما در واقع نشان می­دهد که روش ممیزی توانایی بالایی برای بخاطر سپردن داده های آموزشی دارد و بسیار مستعد بیش برازش[[38]](#footnote-38) است.

**استخراج ویژگی از متن در روش های ممیزی**

در این بخش به بررسی روشهای استخراج ویژگی از متن برای استفاده در دسته بندی کننده های ممیزی می­پردازیم. ویژگی[[39]](#footnote-39) به معنای گواهی کوچکی است که ارتباطی بین آنچه که مشاهده کردیم (همان داده D) و یک کلاس داده C برقرار می­کند. ویژگی، یک تابع با مقدار محدود و حقیقی است. در ادامه چند نمونه از ویژگی های مورد استفاده در روش های ممیزی برای شناسایی موجودیت نامدار را مشاهده می­کنید.

|  |
| --- |
|  |

و داده آموزشی زیر را داریم.

|  |  |  |  |
| --- | --- | --- | --- |
| 4 | 3 | 2 | 1 |
| PERSON | DRUG | LOCATION | LOCATION |
| saw Sue | taking Zantac | in Québec | in Arcadia |

ویژگی یک (بنفش) سه شرط دارد. اولا کلاس داده LOCATION باشد، دوما کلمه قبلی «in» باشد و خود کلمه با حرف بزرگ شروع شده باشد. هر سه این ویژگی ها در مورد داده های آموزشی دو ستون سمت چپ (1و2) صادق هستند. ولی در مورد دو ستون سمت راست (3و4) درست نیستند. ویژگی دوم می­گوید کلاس داده LOCATION است و کلمه، درون خود کاراکترهای لاتین دارد. این ویژگی تنها در مورد ستون 2 درست است. و سرانجام ویژگی سوم می­گوید، کلاس داده DRUG است و حرف انتهایی «c» باشد. این ویژگی تنها ستون 3 را انتخاب می­کند. در هنگام بررسی ویژگی همیشه فرض می­کنیم که یک محل مشخص در متن داریم که در آن نقطه در حال بررسی هستیم. مثلا در زمان بررسی ستون سوم فرض کردیم که کلمه در حال بررسی «Zantac» است (w=Zantac) و کلمه قبل نیز «taking» است (w-1=taking).

در این مرحله، مدل ممیزی به هر ویژگی یک وزن اختصاص می­دهد. وزن نیز یک عدد حقیقی است. مقدار مثبت برای یک ویژگی نشان می­دهد که آن ویژگی در مورد داده آموزشی صادق بوده است و چیزی است که در متن واقعی نیز رخ داده است. به عنوان نمونه ویژگی اول چون در مورد دو ستون 1و2 صادق است، ممکن است وزنی برابر با 0.9 بگیرد. در نقطه مقابل مقدار منفی برای یک ویژگی، نشان می­دهد که احتمال وقوع آن ویژگی پایین است و آن ویژگی در داده آموزشی صحیح نبوده است. به عنوان نمونه به ویژگی زیر توجه کنید.

این ویژگی به داده ستون اول کلاس DRUG را نسبت می­دهد که غلط است و ما میخواهیم نشان دهیم که احتمالا این ویژگی غلط است و به آن وزن -0.6 را نسبت می­دهیم.

برای محاسبه وزن ویژگی ها، از دو مقدار شمارش تجربی و مدل امید ریاضی[[40]](#footnote-40) ویژگی استفاده می­کنیم. که با دو معادله زیر تعریف می­شوند.

در معدله اول که مربوط به شمارش تجربی است، به سادگی تنها روی داده های آموزشی حرکت کرده و تعداد دفعاتی که ویژگی مورد نظر در آنها درست است را می­شماریم. در روش دوم، توزیع احتمال بر روی جفت داده و کلاس را داریم. با استفاده از آن احتمال تمام جفت کلاس و داده ها را در نظر گرفته و مقدار امید ریاضی ویژگی را محاسبه می­کنیم.

به شکل خاص در کاربرد پردازش متن، ویژگی های تعریف شده دارای دو خاصیت زیر هستند:

1. یک تابع شاخص با مقادیر بله یا نه بر روی خواص داده ورودی.
2. یک کلاس داده

یک ویژگی دلخواه در پردازش متن به صورت زیر نمایش داده می­شود.

یک ویژگی دلخواه بر روی کلاس c و داده d با یک مسند تطبیقی[[41]](#footnote-41) با نام و یک کلاس نمایش داده می­شود. مسند تطبیقی مانند اینکه انتهای کلمه حرف «c» دارد یا کلمه قبلی «in» است یا نه؟ و کلاس داده مانند PERSON یا LOCATION است. مقدار خروجی ویژگی به شکل کلی می­تواند یک عدد حقیقی باشد ولی اینجا ویژگی های ما مقدار یک را در صورت تطبیق با داده و کلاس و در غیر این صورت مقدار صفر را بازمی­گردانند. لذا می­توان گفت هر ویژگی، زیر مجموعه ای از داده را تعیین کرده و یک برچسب برای آن پیشنهاد می­دهد. در مدلهای مبتنی بر ویژگی، تصمیمی که برای یک داده در یک نقطه از متن اتخاذ می­گردد، تنها بر مبنای ویژگی هایی است که در آن نقطه فعال (مقدار یک) هستند.

در ادامه به برخی از موارد استفاده از ویژگی های تعریف شده در پردازش متن می­پردازیم.

1. Information extraction (IE) [↑](#footnote-ref-1)
2. Named entity recognition (NER) [↑](#footnote-ref-2)
3. Part of speech tags (POS) [↑](#footnote-ref-3)
4. Clear factual information [↑](#footnote-ref-4)
5. Entity [↑](#footnote-ref-5)
6. Sentiment analysis (opinion mining) [↑](#footnote-ref-6)
7. Precision recall [↑](#footnote-ref-7)
8. Tagger [↑](#footnote-ref-8)
9. MUC score [↑](#footnote-ref-9)
10. F1 measure [↑](#footnote-ref-10)
11. Sequential model [↑](#footnote-ref-11)
12. Character sub sequences [↑](#footnote-ref-12)
13. Word shape sequences [↑](#footnote-ref-13)
14. Maximum entropy sequence model [↑](#footnote-ref-14)
15. Semantic role labeling [↑](#footnote-ref-15)
16. Tokenization [↑](#footnote-ref-16)
17. Maximum entropy Markov models [↑](#footnote-ref-17)
18. Greedy [↑](#footnote-ref-18)
19. Beam inference [↑](#footnote-ref-19)
20. Viterbi inference [↑](#footnote-ref-20)
21. Dynamic programming [↑](#footnote-ref-21)
22. Conditional random fields [↑](#footnote-ref-22)
23. Conditional likelihood [↑](#footnote-ref-23)
24. State of the art [↑](#footnote-ref-24)
25. Discriminative estimation [↑](#footnote-ref-25)
26. Generative models [↑](#footnote-ref-26)
27. Naïve Bayes [↑](#footnote-ref-27)
28. Language models [↑](#footnote-ref-28)
29. Conditionals discriminative models [↑](#footnote-ref-29)
30. Joint distribution [↑](#footnote-ref-30)
31. N-gram model [↑](#footnote-ref-31)
32. Hidden Markov model [↑](#footnote-ref-32)
33. Logistic regression [↑](#footnote-ref-33)
34. Support vector machine [↑](#footnote-ref-34)
35. Probabilistic graphical models [↑](#footnote-ref-35)
36. Word sense disambiguation [↑](#footnote-ref-36)
37. Smoothing [↑](#footnote-ref-37)
38. Overfitting [↑](#footnote-ref-38)
39. Feature [↑](#footnote-ref-39)
40. Expected value [↑](#footnote-ref-40)
41. Matching predicate [↑](#footnote-ref-41)